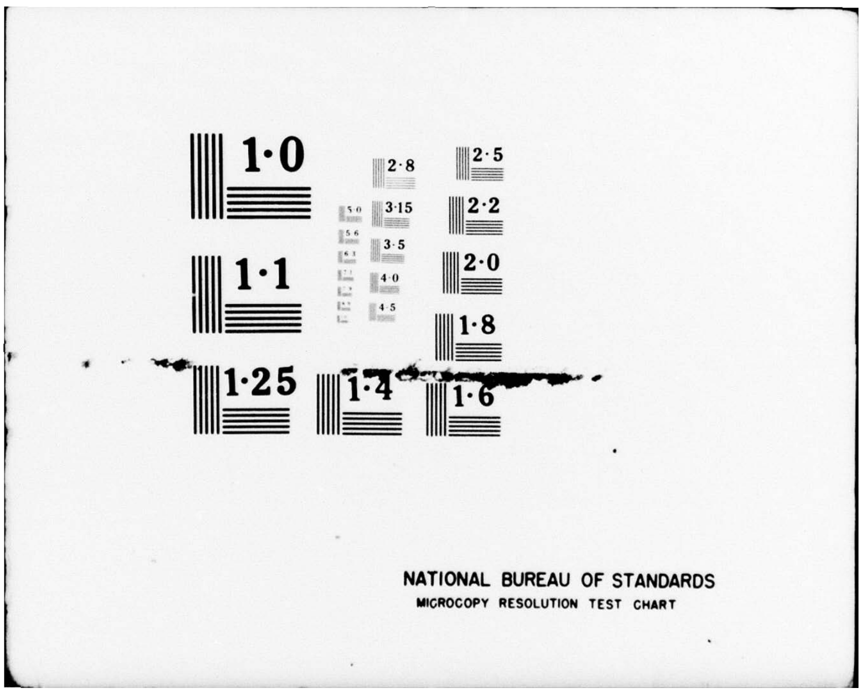
AD-A044 817 MARYLAND UNIV COLLEGE PARK DEPT OF COMPUTER SCIENCE F/G 12/1
SENSITIVITY COEFFICIENTS FOR THE EFFECTS OF ERRORS IN THE INDEP--ETC(U)
SEP 77 G W STEWART N00014-76-C-0391

UNCLASSIFIED TR-571 NL

1 OF 1
ADA
044817

END
DATE
FILMED
10-77
DDC


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·5
2·8
3·15
3·5
4·0
4·5
NATIONAL BUREAU OF STANDARDS
MICROCOPY RESOLUTION TEST CHART

(12)
B.S.

COMPUTER SCIENCE TECHNICAL REPORT SERIES

UNIVERSITY OF MARYLAND
COLLEGE PARK, MARYLAND
20742

14 TR-571

12 19 p.

9 Technical Report, TR-571

ONR-N00014-76-C-0391-571 15

11 Sep[redacted]977

6 **Sensitivity Coefficients for the Effects of Errors in the Independent Variables in a Linear Regression.**

by

10 G. W. Stewart

Abstract

This paper is concerned with errors in the observed values of the independent variables of a linear regression. We propose sensitivity coefficients to measure the effects of these errors and show that they can easily be computed from quantities ordinarily calculated in performing the regression.

This work was supported in part by the Office of Naval Research under Contract No. N00014-76-C-0391.

409 022

Sensitivity Coefficients for the Effects of Errors in the Independent Variables in a Linear Regression

G. W. Stewart

1. Introduction

In this paper we shall be concerned with the regression problem

$$\text{minimize } \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2,$$

where $\mathbf{X}$ is an $n \times p$ matrix of rank p, $\mathbf{y}$ is an n-vector, and $\|\cdot\|$ denotes the usual Euclidean vector norm.* The problem has the unique solution

$$\boldsymbol{\beta} = \mathbf{X}^{\dagger}\mathbf{y} \tag{1.1}$$

where $\mathbf{X}^{\dagger} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{T}$ is the pseudo-inverse of $\mathbf{X}$.

Although classical regression theory concerns itself with the statistical analysis of errors in the vector y, it frequently happens that the design matrix $\mathbf{X}$ is itself contaminated with errors, so that one is effectively working with a perturbed matrix $\mathbf{X} + \mathbf{E}$. For example, the columns of $\mathbf{X}$ may be measured by means of some instrument for which the originator of the problem can only give crude error estimates.
In this case the data analyst is faced with the problem of deciding when the effects of the errors can be ignored. The problem is especially

*In the sequel $\|\cdot\|$ will also denote the spectral matrix norm defined by $\|\mathbf{A}\| = \sup\{\|\mathbf{A}\mathbf{x}\| : \|\mathbf{x}\|=1\}$. See [3] for details.

critical in general purpose regression routines, where it is desirable to provide the user with a set of easily interpretable numbers that indicate the magnitude of the effects of the errors.

A partial solution is provided by the perturbation theory for the regression problem (for a survey of this theory see [4]). This theory bounds perturbations in β in terms of $\|E\|$ and of the "condition number" $\kappa = \|X\|\|X^{+}\|$. Although the results of this theory shed considerable light on the behavior of regression problems under perturbations in X, they are unsatisfactory in practice for two reasons. First they bound only the norm of the perturbation in β, so that a large perturbation in one component can conceal the fact that the others have small perturbations. More important, they are not scale invariant; changing the scale of the columns of X will change κ, even though the statistical problem is essentially unaltered. This phenomenon makes the results quite difficult to interpret.

Taking a different approach, Beaton, Rubin, and Barone [1] have derived measures of sensitivity that to some extent answer the above objections. However, it is assumed that the errors are unbiased and n is large. Swindel and Bauer [5] have derived a useful bound for the relative bias in β, which measures the relative effects of perturbations in X compared to the usual statistical errors in y. In this paper we shall derive coefficients γ_{ij} that measure the sensitivity of β_i to changes in column j of X. Specifically γ_{ij} is the norm of the Frechet

derivative of β_i regarded as a function of the j-th column of X. If ε is the norm of the perturbation of the i-th column of X, then $\gamma_{ij}\varepsilon$ will be an asymptotic bound on the perturbation induced in β_j.*

2. Derivation of the Coefficients

Although it is in principle possible to calculate the required derivatives directly from the normal equations $(X'X)\beta = X'y$, we prefer to approach the problem through a first order perturbation theorem that is useful in its own right.

Theorem 2.1. In the notation of the last section, let β be the solution of the regression problem (1.1), and let $r = y - X\beta$ be the corresponding residual vector. Let E be an $n \times p$ matrix. If

$$\|X^{+}E\| < 1 , \tag{2.1}$$

then $X + E$ has rank p so that there is a unique solution $\tilde{\beta}$ of the problem

$$\text{minimize } \|y - (X+E)\tilde{\beta}\|^{2} .$$

Moreover, as E approaches zero

$$\tilde{\beta} = \beta - X E\beta + (X'X)^{-1}E^{T}r + O(\|E\|^{2}) .$$

Proof. For a proof that (2.1) implies that $X + E$ has full rank, see [4]. This implies that for all sufficiently small E, $\tilde{\beta}$ exists and is given by

*Here, and throughout this note, the term asymptotic refers to behavior for small ε, not large n.

$$\tilde{\beta} = [(X+E)'(X+E)]^{-1}(X+E)y .$$

Now it is well known [3] that if P is sufficiently small, then $I + P$ is nonsingular and

$$(I+P)^{-1} = I - P + O(\|P\|^2) .$$

It follows that

$$\begin{aligned}
[(X+E)'(X+E)]^{-1} &= (X'X+X'E+E'X+E'E)^{-1} \\
&= \{X'X[I+(X'X)^{-1}(X'E+E'X+E'E)]\}^{-1} \\
&= [I-(X'X)^{-1}(X'E+E'X)](X'X)^{-1} + O(\|E\|^2) .
\end{aligned}$$

Hence

$$\begin{aligned}
\tilde{\beta} &= (X'X)^{-1}X'y + (X'X)^{-1}E'y - (X'X)^{-1}X'E(X'X)^{-1}X'y \\
&\quad - (X'X)^{-1}E'X(X'X)^{-1}X'y + O(\|E\|^2) \\
&= \beta - X^{\dagger}E\beta + (X'X)^{-1}E'(y-X\beta) \\
&= \beta - X^{\dagger}E\beta + (X'X)^{-1}E^{T}r . \square
\end{aligned}$$

Theorem 2.1 immediately gives an expression for the Frechet derivative of β_i regarded as a function of the j-th column of X.

<u>Corollary</u> 2.2. Let $\beta_i = f_{ij}(x_j)$, where x_j denotes the j-th column of X. Then

$$Df_{ij} = -\beta_j e_i^T X^{\dagger} + e_i'(X'X)^{-1}e_j r^T . \tag{2.2}$$

Proof. The Frechet derivative Df_{ij} is the unique row vector satisfying

$$f_{ij}(x_j+e) = \beta_i + Df_{ij}e + O(\|e\|^2) . \tag{2.3}$$

Let e_j denote the j-th unit vector. Then a perturbation e in x_j amounts to adding to X the matrix $E = ee_j^T$. Hence from Theorem 2.1

$$\begin{aligned} f_{ij}(x_j+e) &= e_i'[\beta - X^{\dagger}(ee_j')\beta + (X'X)^{-1}(ee_j')'r] + O(\|e\|^2) \\ &= \beta_i - \beta_j e_i'X^{\dagger} e + e_i'(X'X)^{-1}e_j e'r + O(\|e\|^2) \\ &= \beta_i - (\beta_j e_i'X^{\dagger} - e_i'(X'X)^{-1}e_j r')e + O(\|e\|^2) , \end{aligned}$$

which shows that f_{ij}' defined by (2.2) satisfies (2.3).□

Corollary 2.3. Let $C = (X'X)^{-1}$. Then

$$\|Df_{ij}\| = \gamma_{ij} \equiv \sqrt{|\beta_j|^2 c_{ii} + \|r\|^2 c_{ij}^2} .$$

Proof. The vector r is orthogonal to the column space of X, which is the same as the row space of $X^{\dagger}$. Hence $e_i^T X^{\dagger}$ and r' are orthogonal, and

$$\|Df_{ij}\|^2 = |\beta_j|^2 \|e_i X^{\dagger}\|^2 + c_{ij}^2 \|r\|^2 .$$

The proof will be complete if we can show that $\|e_i'X^{\dagger}\|^2 = c_{ii}$. But

$$\|e_i'X^{\dagger}\|^2 = \|e_i'(X'X)^{-1}X'\| = e_i'(X'X)^{-1}X'X(X'X)^{-1}e_i = e_i'(X'X)^{-1}e_i = c_{ii} .□$$

3. Applications

It should be observed that the numbers γ_{ij} can be calculated from the quantities that are usually computed in the course of the regression. For example, if sweep techniques (e.g. see [2]) have been used to solve the problem, the numbers c_{ij} will be available, as will $\|r\|^2$, since it is simply the residual sum of squares.

The results are readily interpretable. If a bound ε_j on the norm of the error in the j-th column is available, then the error δ_i induced in β_i is asymptotically bounded by $\gamma_{ij}\varepsilon_j$. If this number is too large, the problem requires further study. In interpreting the results it should always be borne in mind that $\gamma_{ij}\varepsilon_j$ is a first order bound. A large value is a signal that something may be wrong, but if ε_j is so large that the first order approximation is not applicable, then the difficulties may turn out to be illusory.

A particularly attractive feature of the asymptotic bounds is that, since they deal with indivual components of β, their interpretation is independent of the scaling of the columns of X. This is particularly apparent if the bounds are cast in terms of relative error in the form

$$\frac{|\delta_i|}{|\beta_i|} \lesssim \left(\frac{\gamma_{ij}\|x_j\|}{\|\beta_i\|}\right)\frac{\varepsilon_j}{\|x_j\|} .$$

Now

$$(3.1) \qquad \frac{\gamma_{ij}\|x_j\|}{|\beta_i|} = \left[(|\beta_j|^2\|x_j\|^2)\left(\frac{c_{ii}}{|\beta_i|^2}\right) + (\|r\|^2)\left(\frac{|c_{ij}|^2\|x_j\|^2}{|\beta_i|^2 \quad \|}\right)\right]^{\frac{1}{2}} ,$$

and each parenthesized term in the right hand side of (3.1) is easily seen to be invariant under scaling of the columns of $\underset{\sim}{X}$. Indeed, in some applications where the β_i's are known to be bounded away from zero it may be more appropriate to report $\gamma_{ij}\|\underset{\sim}{x}_j\|/|\beta_i|$ than γ_{ij}.

In deriving our bounds, we have used the Cauchy-Schwarz inequality $\|\underset{\sim}{x}^*\underset{\sim}{y}\| \leq \|\underset{\sim}{x}\|\|\underset{\sim}{y}\|$, an inequality which is usually pessimistic, since it must account for the worst case where $\underset{\sim}{x}$ and $\underset{\sim}{y}$ are dependent. If we are willing to assume more about the perturbation $\underset{\sim}{e}$ in $\underset{\sim}{x}_j$, then we may be able to say more. For example, we have the following consequence of Corollary 2.3.

Corollary 3.1. Let $\underset{\sim}{e} \in N(0,\sigma^2\underset{\sim}{I})$. Then $\underset{\sim}{D}f_{ij}\underset{\sim}{e}$ is normally distributed with mean zero and standard deviation $\gamma_{ij}\sigma$.

References

1. A. E. Beaton, D. B. Rubin, and J. L. Barne, The acceptability of regression solutions: another look at computational accuracy, J. Amer. Stat. Assoc. 71 (1976) 158-168.

2. A. P. Dempster, Elements of Continuous Multivariate Analysis, Addison-Wesley, Reading, Massachusetts (1969).

3. G. W. Stewart, Introduction to Matrix Computations, Academic Press, New York (1973).

4. ____________, On the perturbation of pseudo-inverses, projections, and linear least squares problems, to appear SIAM Rev.

5. B. F. Swindel and D. R. Bower, Rounding errors in the independent variables in a general linear model, Technometrics 14 (1972) 215-218.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE — READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER ONR-N00014-76-C-0391-571	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SENSITIVITY COEFFICIENTS FOR THE EFFECTS OF ERRORS IN THE INDEPENDENT VARIABLES IN A LINEAR REGRESSION		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER TR-571
7. AUTHOR(s) G. W. Stewart		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0391
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Computer Science University of Maryland Colege Park, MD 20742		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Mathematics Branch Office of Naval Research Arlington, VA 22217		12. REPORT DATE September 1977
		13. NUMBER OF PAGES 9
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

regression
least squares
errors in variables
sensitivity

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper is concerned with errors in the observed values of the independent variables of a linear regression. We propose sensitivity coefficients to measure the effects of these errors and show that they can easily be computed from quantities ordinarily calculated in performing the regression.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)